# From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI

Year: 2022 | Citations: 536 | Authors: Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters

---

## Abstract

The rising popularity of explainable artificial intelligence (XAI) to understand high-performing black boxes raised the question of how to evaluate explanations of machine learning (ML) models. While interpretability and explainability are often presented as a subjectively validated binary property, we consider it a multi-faceted concept. We identify 12 conceptual properties, such as Compactness and Correctness, that should be evaluated for comprehensively assessing the quality of an explanation. Our so-called Co-12 properties serve as categorization scheme for systematically reviewing the evaluation practices of more than 300 papers published in the past 7 years at major AI and ML conferences that introduce an XAI method. We find that one in three papers evaluate exclusively with anecdotal evidence, and one in five papers evaluate with users. This survey also contributes to the call for objective, quantifiable evaluation methods by presenting an extensive overview of quantitative XAI evaluation methods. Our systematic collection of evaluation methods provides researchers and practitioners with concrete tools to thoroughly validate, benchmark, and compare new and existing XAI methods. The Co-12 categorization scheme and our identified evaluation methods open up opportunities to include quantitative metrics as optimization criteria during model training to optimize for accuracy and interpretability simultaneously.